The International Journal of Digital Curation
Issue 2, Volume 1 | 2007

Data Curation Standards and Social Science Occupational Information Resources

Paul Lambert, Vernon Gayle,
Applied Social Science
University of Stirling

Larry Tan, Ken Turner,
Computing Science and Mathematics,
University of Stirling

Richard Sinnott,
National e-Science Centre,
University of Glasgow

Ken Prandy,
Cardiff School of Social Science,
University of Cardiff

June 2007

Abstract

Occupational information resources - data about the characteristics of different occupational positions - are widely used in the social sciences, across a range of disciplines and international contexts. They are available in many formats, most often constituting small electronic files that are made freely downloadable from academic web pages. However there are several challenges associated with how occupational information resources are distributed to, and exploited by, social researchers. In this paper we describe features of occupational information resources, and indicate the role digital curation can play in exploiting them. We report upon the strategies used in the GEODE research project (Grid Enabled Occupational Data Environment[1]). This project attempts to develop long-term standards for the distribution of occupational information resources, by providing a standardized framework-based electronic depository for occupational information resources, and by providing a data indexing service, based on e-Science middleware, which collates occupational information resources and makes them readily accessible to non-specialist social scientists.

[1] http://www.geode.stir.ac.uk/

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Background

The analysis of occupational positions is a staple component of social science research. In sociology in particular, the connection between occupational and social structures is deeply ingrained in an array of theoretical accounts (Wright, 2005). Equally, occupational analyses can be found across numerous other research disciplines, as seen for instance in traditions studying labor incomes in economics (Routh, 1980); the health impacts of occupational inequalities in epidemiology (Arber, 1997); the evolution of social and occupational associations in social history (van Leeuwen, Maas, & Miles, 2005); and studies of interpersonal relations in social psychology (Burchell et al., 1999).

Occupational information', as defined here, is used by social scientists to make sense of 'source occupational data'. Occupational information constitutes data about the characteristics of different occupational positions. This may be linked to 'source' data (such as survey questionnaire responses) on the particular occupational positions held by the subjects of analysis. In almost all examples, researchers wish to use occupational information in order to summarize data about their subjects' occupational circumstances in a substantively meaningful but parsimonious way. As a well known illustration, sociologists often wish to use occupational information in order to classify individuals into 'social class' groups on the basis of their current occupation. Indeed, deriving occupation-based social classifications such as social class groups is a particularly well-developed aspect of occupational information provision. Nevertheless, it is important to appreciate that social scientists make use of many other occupational information resources, another example being statistical databases on occupational circumstances, say on average incomes by occupational groups (McKnight & Elias, 1997), or the proportion of women in different occupations (Hakim, 1998).

Existing Occupational Information Resources

The common interests shared by many social scientists to summarize source occupational data has not been matched by a shared exploitation of the same methods for processing occupational information. In fact there is little agreement over which occupational information resources (OIRs) should be favoured from an array of options. There has also been little consistency in how different researchers store, disseminate, and process, different OIRs.

To indicate the scale of the topic, it can reasonably be claimed that thousands of different occupational information resources are available to social scientists. Table 1 below describes an illustrative selection, though there are numerous further resources.

Format	Index units [#]	Output	Documentation	# files / size	ISI citations[1]
1 CAMSIS derivation matrices, http://www.camsis.stir.ac.uk/ [2001; >10 revisions]					
SPSS / plain text	*OUG; e.s; gender [2000]*	*Scale scores*	√√	*200 / 100kb*	*5*
2.1 ISEI tools, http://home.fsw.vu.nl/~ganzeboom/pisa/ [1992; est. 5 revisions]					
SPSS	*OUG [533]*	*Scale scores*	√√	*20 / 50kb*	*61 est.*
2.2 ISEI tools at IDEAS-REPEC, http://ideas.repec.org/c/boc/bocode/s425802.html [2002; 1 revision]					
Stata	*OUG [533]*	*Scale scores*	√√√	*20 / 50kb*	*31 est.*
3.1 ESeC matrices, http://www.iser.essex.ac.uk/esec/ [2005; est. 3 revisions]					
MS-Excel; SPSS	*OUG; e.s. [4000]*	*Social class*	√√	*3 / 100kb*	*0*
3.2 NS-SEC matrices, http://www.statistics.gov.uk/methods_quality/ns_sec/ [2001; no revisions]					
MS-Excel; pdf	*OUG; e.s. [3000]*	*Social class*	√√	*6 / 200kb*	*24*
4. Hakim (1998) gender segregation codes [1998; no revisions]					
Book	*OUG [400]*	*%female per job*	√	*2 / n.a.*	*38 est.*
5. HISCO occupational labels and codes, http://historyofwork.iisg.nl/ [2003; no revisions]					
Book; web html	*OUG [500]*	*OUG content*	√	*1 / n.a.*	*13*
6. O*NET database, http://www.occupationalclassifications-titles.net/ [2001; >10 revisions]					
Web html	*OUG [950]*	*OUG content*	√	*100 / n.a.*	*7*
7. Wright (1985) class scheme classification instructions [1985; 1 revision]					
Book	*Job content data*	*Social class*	√	*2*	*100 est.*
8. IPUMS unit group labels, http://international.ipums.org/international/ [2000; > 10 revisions]					
Web html	*OUG [500]*	*Text labels*	√√	*68 / 50kb*	*n.a.*

- 'Index units [#]': type of index, and average number of distinct index units, in resource.
- '# files / size': approximate number of distinct files, and typical size of each, in resource.
- [] : year of first publication, and number of subsequent updates to contents.
- OUG = occupational unit group; e.s. = 'employment status'.
- Documentation: √ / √√ / √√√ = brief natural language / extended natural language / metadata.

Table 1 Selected occupational information resources in the social sciences.

Table 1 also illustrates that there is little coordination between alternative occupational information resources. Perhaps the only common factor unifying OIRs is that most resources feature some sort of definition of occupational positions into 'index units' (see discussion below). However, different occupational information resources use different index unit definitions. Moreover, different OIRs tend to supply information in different data formats, and to provide different levels of documentation on those formats. Perhaps most significantly, there are very few examples of OIRs which contain standardized metadata describing their origin and content. Most resources do feature explanatory notes and natural language documentation, but the systematic contribution of metadata is rare (the only exceptions known to the authors are the ISEI classification tools stored at the IDEAS-REPEC depository - 2.2 in Table 1 – which conform to the metadata requirements of the IDEAS system).

Table 1 illustrates the abundance of occupational information resources available to social scientists, which can be explained by a number of trends in social science research practice. One involves interest in studying occupations across different countries in internationally and/or historically comparative analyses. Although some

OIRs are specific to certain countries and time periods (for instance, Hakim's 1998 gender segregation codes apply to Britain over the 1990's), most feature a significant international and/or historical coverage. For instance, the CAMSIS, HISCO and IPUMS resources cited in Table 1 provide different resources for, respectively, 30, 12 and 20 different nations. These three projects also publish different OIRs relevant to different historical periods – the resources of the CAMSIS and IPUMS projects span the nineteenth century to the present day, whilst those of HISCO span the sixteenth to the twentieth century. The ISEI and ESeC tools cited in Table 1 also provide international OIRs, though these tools are standardized so that the same resources apply to different nations (cf. Ganzeboom & Treiman, 2003).

The abundance of OIRs can also be explained by wider motivations to provide different types of occupational information appropriate to the range of social science disciplines and interests which make use of occupational data. In Table 1, resources 1-3.2 and 7 all offer facilities to calculate other occupation-based social classifications. However, resources 4-6 offer more specific, descriptive data on the nature of occupational units. The O*NET database cited in Table 1, for example, is designed to provide accessible descriptions of contemporary occupations in the United States, for the benefit of job seekers in that country.

A further explanation for the proliferation of OIRs involves upgrades and revisions to previously published resources. Sometimes such upgrades occur in response to major consultations concerned with the changing meanings of occupational positions over time, for instance through updates to dictionary definitions of occupational titles (e.g. ILO, 1969, 1990; ONS, 2000; OPCS, 1980, 1990), or in revisions in social class classifications (e.g. Rose & Pevalin, 2003). Upgrades and revisions may also occur when the publishers of empirically derived occupational information access new or revised empirical data on which to base their calculations (e.g. Ganzeboom & Treiman, 1996). More prosaically, revisions may also arise when the individuals publishing occupational information decide to change some of the values in their data, perhaps for theoretical reasons (a well known example being Wright's 1985 revision to his social class scheme).

Current Practices and Problems in Occupational Analysis

The abundance of OIRs leads directly to a number of challenges for the many non-specialist social scientists who wish to access and exploit occupational information[2].

Limited coordination in the distribution of existing occupational information resources reduces accessibility for potential users of occupational data. Under current provisions, social scientists who wish to analyze occupational records are usually required to undertake two relatively challenging tasks. First, they must navigate across diverse internet locations and choose between the numerous available OIRs. Second, they must implement a connection between the relevant published resource(s), and the occupational records in their own database. Since OIRs tend to be released in a variety of different electronic formats, this means that social scientists may be required to use

[2] Moreover, the challenges resulting from the abundance of OIRs may be expected to expand over time. For instance, the dissemination of OIRs through internet sites has grown rapidly in the last decade. It has also been observed that recent decades have seen increased access to large-scale data resources which in turn permit the generation of new occupational information (e.g. Goldthorpe, 2005).

multiple software packages and to undertake a variety of tasks to manipulate the occupational data in order to allow its connection with their own requirements. Both of these processes have hitherto proved difficult for many social researchers.

One revealing insight into current practices in occupational analysis is the data conveyed in the last column of Table 1, intended to be indicative of the uptake of published occupational information resources[3]. A well disciplined model of social science investigation would see most researchers using published occupational information in a consistent and well documented way (Goldthorpe, 2005). However Table 1 suggests that the practice of occupational analysis is far more 'messy' than this model. Relatively few people, as a proportion of those conducting occupational analysis, appear to utilize these published resources. Indeed, when they do use published resources, Table 1 suggests that users favor those resources which have more limited documentation and simpler formats (such as the text publications 4 and 7). Several previous authors have noted that, by contrast, social science researchers are much more likely to construct their 'own' occupational information, and deploy it in their own idiosyncratic (and undocumented) style (Bechhofer, 1969; Lambert, 2002; Marsh, 1986).

A related theme in the exploitation of occupational information resources has been the persistent failure of attempts by individuals and organizations to assert that certain standardized OIRs should be used in all relevant social science research. Such assertions are especially common with regard to the use of occupational data to derive occupation-based social classifications. Here, various academics, and national and international statistics agencies, have recommended alternative standard classifications which should be used for all relevant analyses. In Britain, for example, the evolution in the recommended methods for classifying occupations to the officially endorsed 'Registrar General's Social Class scheme', used between 1911 and 2002, and latterly its replacement with the 'National Statistics Socio-Economic Classification', have been well documented (Rose & Pevalin, 2003; Szreter, 1984). Nevertheless, such attempts to promote standardizations have been overwhelmingly unsuccessful. Although several textbooks in research methods do urge social scientists to exploit certain OIRs, the actual practice evident from published research is far less consistent (Reid, 1998).

Another handle on current practices can be gained by reviewing examples of published research which exploits occupational data. A variety of standards can be uncovered. As illustrations, analyses by Bihagen and Ohls (2006) and Platt (2005) can be heralded as research which maximises the evaluation of candidate occupational information and the provision of documentation. Bihagen and Ohls' work involved linking survey data with three different occupationally based social classifications by using published index files which are described in the text; the analysis incorporated an evaluation of the relative properties of the three classifications considered. Platt's work involved selecting a single occupationally based social classification for analysis, on the basis of a sequence of explicit decisions about the quality of its documentation

[3] The number of ISI indexed journal articles citing the relevant occupational information resource's documentation in their bibliography (calculated from Web of Knowledge citation statistics, http://www.wok.mimas.ac.uk , in November 2006). The figures for the ISEI tools show an estimated fraction of 92 studies citing the ISEI documentation. The figures for the Hakim and Wright texts are estimates derived from the total number of citations of those books.

and comparability; this decision-making process was discussed in detail in an online appendix. However it should be recognized that the effort and skills involved in these implementations are substantial, and relatively few social scientists have demonstrated the diligence illustrated by these studies.

The analyses of Archer and Francis (2006), Dixon and Paxton (2005) and Modood (2005) may be presented as more problematic examples of occupational research. All implement classifications of occupational positions on the basis of occupational information. However Archer and Francis's shortcoming involves their designation of an occupational class classification based upon their own judgment, which generates a scheme which is not replicable and may not be readily compared with other published analyses. Dixon and Paxton's weakness involves their attempt to synthesize results from a series of occupationally based social class classifications which are not equivalent, without providing details on the implications of alternative schemes or their comparability. Modood's text similarly lacks details on the specification of a simplistic occupation-based social class classification which is used to help explain patterns of educational attainment (p. 300ff). The problems are exacerbated in this instance because, as Modood notes, complexities in respect of the labor market situations of the ethnic groups studied in this analysis are ignored by the occupation-based social classification used. Implicitly, this recognition suggests that there may be stronger relationships between occupational circumstances and the inequalities under study, but that this analysis is not able to reveal them, because of a limitation in the occupational classification used. In each of these examples, the limitations associated with the outputs suggest that the researchers were not in a position comfortably to review a wider range of potentially relevant occupational information resources, nor undertake and document a clearly defined linkage between their data and suitable OIRs.

Strategies for Managing Occupational Data

As noted above, one reaction to the inconsistencies evident in social researchers' use of occupational information resources has been to try to impose standardization on the collection and analysis of occupational data, for instance by enforcing data collectors to code records into a standardized occupational scheme, and by asserting that certain occupational information resources should be preferred over rivals. Attempts at this strategy have been, hitherto, unsuccessful. This certainly stems, in part, from ill-discipline amongst social scientists[4]. However there are also theoretical arguments for rejecting this approach. The attempt at standardization is well represented by 'universal' approaches to occupationally based social classifications (also referred to in methodological texts as approaches of 'measurement equivalence', e.g.Hoffmeyer-Zlotnik & Wolf, 2003). Here it is asserted that occupational structures across different countries and time periods, and between men and women, are broadly stable. This implies that a single occupationally based social classification is adequate for all research investigations – the claim that the same occupational title means the same thing across countries, time periods, and between men and women (Hout & DiPrete, 2006, p.2). However, a universal approach to occupational information

[4] Indeed, almost forty years ago, Bechhofer's review of the use of occupational information in sociology bemoaned the abundance of, and inconsistencies between, occupationally based social classifications, noting that *"..researchers are advised not to add to the already existing plethora of classifications without very good reason"* (Bechhofer, 1969, p.118). However since that recommendation, the number of new classifications has increased steadily.

resources can be demonstrated to be both theoretically and empirically unsatisfactory, since it is prone to neglect patterns of occupational change and contextual differences in occupational experiences (Lambert, Tan, Gayle, Prandy, & Turner, (in press)). On methodological grounds, approaches of 'functional equivalence' – which can allow occupational measures to vary across time or counties – are often presented as more favorable (Hoffmeyer-Zlotnik & Wolf, 2003). Moreover the pluralistic theoretical traditions of social scientists (Wright, 2005) suggest that a universal approach to occupational information is, in practice, unattainable.

A pluralistic approach to managing diverse occupational information resources is therefore more attractive than standarization. Here, researchers' access to alternative resources may be facilitated and encouraged, and standards of explicit documentation and evaluation fostered. One option could be to provide informative textual comparisons of occupational information resources. Several texts have provided focused reviews of selected resources (Hakim, 1998, Annex A). However, there has been no widely accepted systematic summary of all occupational information resources, and prospects for such an undertaking would seem unlikely, given the variety and volume of resources involved.

An alternative is a computer-based facility for describing occupational information. The GEODE Project[5] seeks to provide an online database which collates data on occupational information resources and distributes it across the social science research community. It attempts to develop long-term standards for the distribution of occupational information resources, by providing a standardized framework-based digital depository for occupational information resources, and by providing a data-indexing service which collates occupational information resources and makes them readily accessible to non-specialist social scientists.

A particular feature of the GEODE Project is that it seeks to exploit the capabilities conferred by e-Science computing in making these provisions. This strategy is, in part, driven by an e-Science agenda, since GEODE is a project funded through a specialist research programme, coordinated by the National Centre for eSocial Science[6], which aims to evaluate e-Science capabilities, and to develop capacity in e-Science computing, for social science applications. The development of the GEODE service used a series of conceptualizations which illustrated how e-Science services associated with security, data abstraction and virtual organizations could contribute to user requirements in working with OIRs (Tan, Gayle, Lambert, Sinnott, & Turner, 2006). The concluding section of this article features comments on the contribution of these facilities to the GEODE data service.

[5] http://www.geode.stir.ac.uk/
[6] http://www.ncess.ac.uk/

Curation of Occupational Information Resources

Metadata Requirements

The existing arrangements for the distribution of occupational information resources exhibit a clear shortcoming, namely the absence of consistently structured metadata. It is well recognized that consistent standards of data curation through metadata enable rapid navigation and processing of information resources[7]. This is particularly true in the context of Grid-enabled datasets (Cole, Schurer, Beedham, & Hewitt, 2003). Therefore an objective of the GEODE Project has been to establish a framework for the curation of occupational information resources.

Following earlier recommendations on e-Social Science standards (Cole et al., 2003), and in line with prevailing practices in curation of other social science datasets (Blank & Rasmussen, 2004), GEODE uses a data curation structure based upon the Michigan Data Documentation Initiative (DDI, version 2.1[8]). This standard is attractive because the storage of metadata in a DDI format allows ready integration with the data manipulation processes also catered for in GEODE. The DDI offers a generic set of XML tags which can be used to curate in a consistent manner a large range of social science data. The GEODE Project is concerned with a limited range of metadata statements, those required to curate adequately the small data files typical of occupational information resources. Moreover, as such data files are often updated over time, there is motivation to find DDI-based standards of curation that are relatively quick to implement.

The GEODE Project therefore concentrates upon a prescribed subset of DDI tags, referred to as the 'GEODE-M' metadata standard. A review of existing occupational information resources was undertaken in order to establish which information was most important to generating metadata on the occupational records. Three critical contexts were identified:

Index schemes for source occupational data.

In most social surveys, a textual description of the occupational title and circumstances is taken as the initial source occupational record. This information may be stored as free text. However, more commonly it is translated into an index of occupational positions, usually a location within an 'occupational unit group' (OUG) scheme. In most countries, prescriptive documents are available which show how occupational descriptions may be assigned to numerically standardized occupational schemes, such as OUG systems (ILO, 1990; ONS, 2000), or industrial sector classifications (ONS, 2003). In several cases, computer software is available to allow rapid classification of textual occupational descriptions into numerical OUG locations[9].

---

[7] http://www.dcc.ac.uk/resource/curation-manual/

[8] A revised version of the DDI scheme, version 3.0, was introduced in spring 2007. This version has compatibility with version 2.1., although some programming is required to achieve this. The discussions in this article refer to DDI version 2.1. http://www.icpsr.umich.edu/DDI/

[9] E.g. Computer Assisted Structured Coding Tool, http://www2.warwick.ac.uk/fac/soc/ier/publications/software/cascot/

Three types of index scheme are commonly used to preserve source occupational data. One concerns the classification of occupational titles into an OUG scheme; it is this type of occupational data which has the widest range of occupational information resources associated with it. Another concerns the industrial sector of the occupation. As indicated above, standardized index schemes are widely used for classifying occupational titles and industrial sectors. A third type of index is most usually described as 'employment status', and concerns the ownership of the occupational site and circumstances of the employment contract. Several standardized employment status indexes exist (Elias, 2000), but many statistical agencies and data collectors use bespoke employment status questions. In addition to these more common types of record, many studies also hold additional data on the occupational position held by an individual – examples include the normal time and days of work; as well as aspects of the work process such as the extent of supervision experienced.

In seeking to provide facilities for the curation of occupational information resources and their relation to source occupational data, the GEODE Project takes as its starting point the assumption that source occupational information has been recorded in the format of a published occupational index scheme such as an occupational unit group (OUG) system. This proves to be an important assumption since published occupational index schemes exhibit the idealized features of a 'standard category' <stdCatgry> record within the DDI structure. The declaration of occupational index schemes as standard categories means that connections between occupational information resources, and source occupational data, can in principle be fully leveraged simply by searching for matching combinations of the relevant index scheme(s).

The declaration of a DDI 'standard category' requires reference to further details on each index scheme. In the case of occupational information, the requirement is for published resources which give authoritative information on the nature of different named occupational index schemes. However, the uneven evolution of occupational information resources in the social sciences means that there are several published conflicts between the precise definitions of index schemes. A comprehensive listing of occupational index units would be beneficial in order to allow immediate specifications on the scope of a given standard category. Within GEODE, this is achieved through the manual publication of a listing of occupational index measures[10]. This listing creates a new definition for every new index scheme introduced to the GEODE service.

Context of occupational data.

Occupational information resources are available across a wide array of different 'contexts'. Most frequently, resources are associated with contexts defined by different nations and/or different time periods. Examples of how OIRs relate to different nations and time periods were given above with reference to Table 1 (including through different national resources, and upgrades and revisions to occupational resources over time). However, other social contexts may also be used to delimit the coverage of occupational information resources. For instance, some resources apply only to the occupations of male or female respondents respectively, or only to other particular

[10] at http://www.geode.stir.ac.uk/ougs.html

social groups[11]. Within the DDI scheme, several tags may be used to define the appropriate context of a given occupational information resource, all of which can be suitably located within the 'study information' <stdyInfo> section of the metadata.

Reference unit for occupational analysis.

A third issue in the recording and processing of source occupational data concerns the 'unit of analysis' to which the occupational information is to be applied. A well known debate within sociological literature concerns whether the occupational class of an individual is best understood in terms of their own current occupation (if working), or by incorporating information on previous occupations, or the occupations of household sharers such as a spouse. Although there have been some recommended (but contested) principles for summarizing occupational data (e.g. Erikson, 1984), the permutations associated with the 'reference unit' for occupational data can become complex (such as how adequately to describe a career sequence of occupational positions; or how to merge occupational records from multiple household sharers). The data management tasks involved in such data complexities are substantial and arguably have prevented many researchers from adequately exploiting their source occupational data. For instance, it is argued that most sociological researchers use the more easily implemented 'individual' level occupational measure, despite convincing empirical support for incorporating household level records (Lambert, 2002). Allowing for potentially different reference units is highly attractive. As noted below, the DDI specification of 'variable groups' readily allows data-matching programs to assign multiple linkages between an OIR and several different occupational records, such as are associated with the occupations of different household sharers or different occupations over the course of a career.

GEODE-M Metadata Standard

The GEODE-M customized metadata standard incorporates entries in each of the five component structures of the DDI. These cover a production statement for the metadata itself; statements on the generation of the occupational information resource; statements describing the data file(s) associated with the resource; data describing the content of the data file(s) of the resource; and space for optional additional statements. Segments of the GEODE-M structure are illustrated in Figure 1 below.

[11] For instance, CAMSIS scale scores for male and female occupations (http://www.camsis.stir.ac.uk/), and the HESA occupational unit group scheme for graduate level occupations, http://www.hesa.ac.uk/manuals/05018/05018a04.htm .

```
<codebook>
<docDscr> … <distStmt> <contact email="pl3@stir.ac.uk"> Paul
    Lambert</contact> </distStmt>
    <prodDate date="2006-07-19" >July 19, 2006</prodDate>
                        …  </docDscr>
<stdyDscr> … <titl>CAMSIS scales for the UK using SOC2000</titl>
    <IDNo agency="GEODE">131</IDNo>
    <distrbtr URI="http://www.camsis.stir.ac.uk">Cambridge Social
    Interaction and Stratification Scales website</distrbtr>
     <stdyInfo> <!-- information about the data context -->
     <sumDscr> <timePrd  event="start" >2000</timePrd>
     <nation abbr="GB">United Kingdom</nation>  </sumDscr>
     </stdyInfo>        … </stdyDscr>
<fileDscr id="gb91soc2000.sav"> …
    <fileName id="gb91soc2000.sav">gb91soc2000.sav</fileName>
                        … </fileDscr>
<dataDscr> …
    <varGrp name="indexs" var="soc2000s ukempsts stdempsts" >
    <concept>Index term</concept> ... </varGrp>
    <varGrp name="outcomes" var="MCAMSISs FCAMSISs">
    <concept>Occupational information</concept> </varGrp>
    <var ID="soc2000s" file="gb91soc2000.sav" >
    <stdCatgry uri="http://www.geode.stir.ac.uk/ougs.html#soc2000">
    Standard Occupational Classification 2000</stdCatgry></var>
                        … </dataDscr>
<otherMat> …            … </otherMat>
 </codebook>
```

Figure 1 Example of key DDI XML tags within 'GEODE-M'

The GEODE-M standard is devised in such a way as to minimize the requirements for describing occupational information resources, whilst successfully drawing out the salient identifying features of those occupational information resources which have been reviewed. Figure 1 illustrates the essential contents of a GEODE-M entry, in this example describing a data resource available from the CAMSIS Project webpages[12]. The figure shows that only a handful of information records need be assigned to curate an occupational information resource. These cover a contact name for the supply of metadata; a title statement for the resource itself and data on the location and date of publication of the resource; a specification identifying the file or files being curated; and statements identifying the variables contained within the file, making the crucial allocation of variables into appropriate 'variable groups'.

A critical feature of the DDI standard is the specification of 'variable groups'. These define the nature of the occupational information. They identify all information as either an 'index' measure or an output measure (the same variables could be included in both groups). This separation allows rapid indexing of OIRs according to which index variables are required and which are available. By exploiting the 'standard category' <stdCatgry> statement, index variables may be identified and coordinated, and allocated to appropriate variable groups.

12 http://www.camsis.stir.ac.uk/

An appealing feature of the DDI standard is the repeatable nature of the file description and data description elements. The former allows occupational information resources which supply data simultaneously in more than one data file to be curated as a single record, and resources searched across the range of files[13]. The latter allows multiple variable definitions within each data file to be specified, allowing the possibility of comparable but not identical content being used to describe closely related variables when necessary.

The GEODE-M metadata standard serves to indicate aspects of occupational information resources which allow standard index searching and linkage exercises (see below). It is intended to allow rapid curation of many occupational information resources, since many potential data suppliers from social science backgrounds are unlikely to be sufficiently motivated to spend long periods curating their own data for the benefit of other users. The GEODE-M specification requires as a bare minimum only eleven information statements (underlined in Figure 1). This allows resources to be deposited to the GEODE index service with minimal manual curation. Files may be submitted to the site through an entry portal which features a short form collecting the minimum required GEODE-M statements in a user-friendly manner (Lambert & Tan, 2007).

Nevertheless, additional metadata will improve the quality of data curation and consequently the accessibility of any OIR for index searching and further linkages. The GEODE strategy allows further curation of metadata for any deposited resource by both the original depositor and members of the GEODE Project. Further details may be voluntarily contributed (through editing of the XML record) to extend the curation process. This is not a user-friendly process, but it is available to social scientists with a specialist interest in occupational information files. Instructions on undertaking this process are given (Lambert & Tan, 2007).

Data Curation and Data Management

The innovation associated with the GEODE-M DDI scheme for curating occupational information resources concerns its use to interlink occupational information resources in a Grid-enabled data environment. GEODE uses the OGSA-DAI middleware (Database Access and Integration[14]) to provide a 'Grid' or 'e-Science' framework for these services[15]. Broadly, the GEODE framework involves the connection of a data indexing service (which harvests data and metadata from OIRs supplied from diverse locations and formats) with a data management service, which supports linkages between OIRs and users' source occupational data (a fuller description of the GEODE repository and services architecture is given in Tan et al., 2006).

[13] This scenario is quite common. For example the CAMSIS website disseminates zip archives containing replicated versions of the same data files in different software formats (http://www.camsis.stir.ac.uk/); the PISA website features a series of related data files which all show mechanisms for translating occupational records from different countries into ISCO categories (Ganzeboom, 2007).

[14] http://www.ogsadai.org/

[15] The OGSA-DAI middleware was attractive for this purpose since, as well as supporting the required data indexing and management services, it has also been adopted in other UK e-Science endeavors.

The data indexing service has the characteristics of a data repository. OIRs may be uploaded to the GEODE server for subsequent distribution, but social scientists may also declare the location of their OIRs, without uploading to the GEODE server, whilst still being indexed within the system. This framework allows curated occupational information resources to be connected and exposed to a virtual organization of social science users, who may themselves exploit data management services for searching and linking occupational data. In GEODE, these services may be accessed by non-specialist users through a user-friendly "portal" interface to the databases which uses GridSphere[16]. The GEODE portal has been publicly accessible[17] since January 2007 (Lambert & Tan, 2007).

It is these data management services which offer substantial improvements in the handling of occupational information amongst social science users. These improvements occur in broadly two contexts:

Robust reviews of occupational index schemes.

The previous provision of occupational information resources has required users to search diverse resources for data stored in exactly equivalent index units to those of the user's source occupational data. As we have indicated above, there are numerous occupational index schemes in currency, because published occupational index schemes are regularly revised and updated over time; because alternative schemes are available for alternative contexts such as different countries and time periods; and because occupational index schemes are usually designed in such a way as to incorporate alternative levels of detail on the occupational location[18]. Moreover, the numeric format used for recording locations within occupational index schemes is sometimes inconsistent. In some schemes trailing zeros are used to indicate subgroup membership instead of hierarchical truncation (for example SOC 2000 'major group 5' may also be indicated as '5000'). Equally, some schemes are recorded in text formats and/or with decimal markers in order to distinguish truncation in occupational data. For example, the ISCO-88 codes generated by the CASOC software (Elias, Halstead, & Prandy, 1993) are generated in 'string' format, whilst the UK 1980 classification (OPCS, 1980) is commonly recorded as a 5-digit classification featuring decimals between the third and fourth digit. The practical upshot of the range of alternative occupational index schemes and formats is that most occupational information resources are readily available for a very limited range of index schemes. Researchers have previously been pushed into selecting occupational information on the basis of an exact match in index variable formats. However the GEODE use of standard category statements opens up opportunities to declare relations between different index variables, bridging the gaps generated by formatting inconsistencies.

Rapid implementation of secure file-matching procedures.

The substantial impediment associated with previous applications in occupational research concerned user difficulties in implementing the software-specific linkage between source occupational data and a published occupational information resource.

[16] http://www.gridsphere.org/

[17] from http://www.geode.stir.ac.uk/

[18] For example, in the UK SOC-2000 classification (ONS, 2000), an occupation may be noted as 'unit-group' '5232 Vehicle body builders and repairers', but it could alternatively be recorded as 'minor group' '523 Skilled metal and electrical trades: Vehicle trades', or as the 'major group' '5 Skilled trades occupations'.

Through its exploitation of OGSA-DAI systems, the GEODE service offers a framework for conducting this linkage in an automated but secure way (recognizing that source data is usually highly sensitive). The mechanics of this linkage hinge on identifying the index linking variables available in the source data, a process enabled by the specification of standard categories for occupational index variables.

Conclusions

The contribution of the GEODE Project to the management of occupational information resources in the social sciences has been twofold.

Occupational Information Review and Digital Curation

Firstly, the interrogation of existing internet-based OIRs revealed a need for consistent curation of OIRs. A framework for achieving that curation has been established by developing a system of metadata requirements which are related to the DDI standard. This has lead to the provision of a new information resource, the GEODE portal, which offers an innovative specialized occupational data access service (Lambert & Tan, 2007).

A crucial consideration by the project members was that the total volume of OIRs to be indexed by GEODE would be finite. This motivated the GEODE portal service to incorporate a limited number of manual requirements within the process of curating OIRs. These were accepted on the understanding that these exercises will overwhelmingly be undertaken by members of the GEODE Project. The expectation is that the benefits to the social science community from this input will be considerable (the first coordinated internet service offering occupational information), whereas the manual input required for curating data would not be excessive.

Early usage of the GEODE service has concurred with these intentions. The additional inputs required during the first stage of data indexing have indeed proved minimal. The most significant requirement has been to ensure the updating of the internet page which is used to list 'standard categories' of occupational index schemes[19]. The additional inputs required during the second stage of curating OIRs are more substantial, typically involving around thirty minutes of data entry time by a member of the GEODE Project. However, new resources to the GEODE service have been supplied at low frequency, meaning that this curation requirement is not prohibitive.

e-Science Evaluation

The GEODE Project was designed explicitly to attempt the implementation of its service requirements though the use of e-Science technologies. An encouraging aspect of the GEODE study was that the data indexing framework was demonstrated to work adequately for a themed group of data resources (occupational information resources). This framework, however, was designed in a generic way, and it may be anticipated that the same framework may be readily extended to other groups of themed social science information resources. Geographical and educational aggregate information files are obvious comparators, as are cross-national macro-economic databases and their requirements of linkage to international micro-data[20].

[19] http://www.geode.stir.ac.uk/ougs.html
[20] cf. http://www.mimas.ac.uk/limmd/

However, the GEODE Project may be typical of many investigative e-Science studies insofar as, for many reviewers, it begs the question of whether the implementation of the complex and demanding e-Science middlewares (OGSA-DAI and Gridsphere in this example) is justified. The primary attractions to using these two technologies are that both are increasingly recognized, at least within the UK, as standard tools for e-Science services (standardization of technologies itself being a key aspect of e-Science computing). In addition, both offer the range of data services and online facilities conceived by requirements analyses undertaken during the GEODE Project (Tan et al., 2006). However, both tools required extended specialized programming, and ongoing specialist support, for their implementation with the GEODE Project; for instance, both middlewares were the source of numerous minor software incompatibilities which led to frequent unanticipated delays during their implementation (e.g. Tan et al., 2006).

It is certainly appropriate to question whether other, less demanding technologies may also have been used for similar purposes. Future work may be necessary to answer this question fully (the funded work of the GEODE Project did not incorporate a comprehensive evaluation of what levels of service might have been obtained from other approaches). At the present stage, it is only realistic to discuss the relative contribution of the e-Science approach in comparison to the previous model of one-way internet distribution of occupational information resources. Three benefits of the e-Science strategy are claimed.

The first is the circular observation that the attempt to Grid-enable occupational information resources encouraged a number of approaches to reviewing, managing and distributing OIRs which might not otherwise have been developed – processes which help bring discipline to data in this specialized field of social science research. For instance, had the project not been aware of the stringent metadata requirements beneficial to the data indexing service used through OGSA-DAI, it is less likely that the social scientists involved would have been motivated to develop the DDI-based data curation standard which was exploited. Equally, because the OGSA-DAI framework can support the harvesting of OIRs which are stored on external web servers and may be adjusted over time, the GEODE service was in turn motivated to develop metadata structures which support the depositing of dynamic OIRs (a facility not available from static webpage distributions). Lastly, because the Gridsphere framework is amenable to supporting virtual organizations of its users, the service is readily able to differentiate two groups of users. On the one hand, ‘guest’ users may review OIRs stored at GEODE, and download them or process file linkage connections. On the other hand, named users can deposit their own occupational data, and collaborate with other users in managing that data through the portal. This latter arrangement constitutes a virtual organization which has proved useful to a number of external users of GEODE who are already specialists in working with OIRs.

Another attraction of the OGSA-DAI indexing of occupational information resources is that this framework promotes a close match between the storage and distribution of data resources, and the processing of analytical queries upon them. A relevant example concerns the indexing of OIRs in terms of occupational index units. As described in section 2, the OGSA-DAI indexing is amenable to reviewing a range of OIRs to search for loosely connected index units and ultimately to promote connections between different OIRs which may not otherwise have been exploited.

A third benefit of the e-Science implementation of the GEODE data service is the capacity to accommodate differential security certifications. Security concerns are a major challenge for one important service provided by GEODE, namely the linkage between OIRs and users' private social science micro-data. The GEODE portal is able to support a JAVA-based application which undertakes this matching process without any security risk to the users' micro-data (the micro-data never leaves the user's machine). This is at present a unique service to the GEODE Project, and one which has been positively evaluated by the social science users who have so far exploited it[21].

Acknowledgements

This research is supported by an ESRC 'Small Grant in e-Social Science', RES-149-25-1015.

References

Arber, S. (1997). Comparing inequalities in women's and men's health: Britain in the 1990s. *Social Science and Medicine, 44*(6), 773-787.

Archer, L., & Francis, B. (2006). Challenging classes? Exploring the role of social class within the identities and achievement of British Chinese pupils. *Sociology, 40*(1), 29-49.

Bechhofer, F. (1969). Occupations. In M. Stacey (Ed.), *Comparability in social research* (pp. 94-122). London: Heinemann (in association with British Sociological Association / Social Science Research Council).

Bihagen, E., & Ohls, M. (2006). The glass ceiling - where is it? Women's and men's career prospects in the private vs. the public sector in Sweden 1979-2000. *Sociological Review, 54*(1), 20-47.

Blank, G., & Rasmussen, K. B. (2004). The data documentation initiative: The value and significance of a worldwide standard. *Social Science Computer Review, 22*(3), 307-318.

Burchell, B., Day, D., Hudson, M., Lapido, D., Mankelow, R., Reed, H., et al. (1999). *Job insecurity and work intensification: Flexibility and the changing boundaries of work*. York: York Publishing / Joseph Rowntree Foundation.

Cole, K., Schurer, K., Beedham, H., & Hewitt, T. (2003). *Grid enabling quantitative social science datasets - A scoping study*. Swindon, UK: ESRC / JISC.

[21] At time of publication the number of social scientists to have used the GEODE portal is low, reflecting that aspects of the service are still in development. Nevertheless feedback from those who have used the service (received by email, through an online form, and through questionnaires distributed during the first project workshop held in January 2007) has, thus far, been overwhelmingly positive, with most interest directed towards the data matching facility. For example, one social scientist who had been introduced to the service reported that she 'strongly agreed' with its rationale and that it was 'extremely likely' that she would use GEODE in the future to help with linking her own data with OIRs – an encouraging response from the target audience of this service.

Dixon, M., & Paxton, W. (2005). The State of the nation: An audit of social injustice in the UK. In N. Pearce & W. Paxton (Eds.), *Social justice: Building a fairer Britain* (pp. 21-61). London: Politicos, with IPPR.

Elias, P. (2000). Status in employment: A world survey of practices and problems. *Bulletin of Labour Statistics*, 1-19.

Elias, P., Halstead, K., & Prandy, K. (1993). *Computer assisted standard occupational classification*. London: HMSO.

Erikson, R. (1984). Social class of men, women and families. *Sociology - The Journal of the British Sociological Association, 18*(4), 500-514.

Ganzeboom, H. B. G. (2007). Tools for deriving status measures from ISKO-88 and ISCO-68. Retrieved June 1, 2007, from http://home.fsw.vu.nl/~ganzeboom/PISA/

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations. *Social Sciences Research, 25*, 201-235.

Ganzeboom, H. B. G., & Treiman, D. J. (2003). Three internationally standardised measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnick & C. Wolf (Eds.), *Advances in cross-national comparison. A European working book for demographic and socio-economic variables* (pp. 159-193). New York: Kluwer Academic Press.

Goldthorpe, J. H. (2005). Progress in sociology: The case of social mobility research. In S. Svallfors (Ed.), *Analyzing inequality: Life chances and social mobility in comparative perspective*. Stanford: Stanford University Press.

Hakim, C. (1998). *Social change and innovation in the labour market : Evidence from the Census SARs on occupational segregation and labour mobility, part-time work and student jobs, homework and self-employment*. Oxford: Oxford University Press.

Hoffmeyer-Zlotnik, J. H. P., & Wolf, C. (Eds.). (2003). *Advances in cross-national comparison: A European working book for demographic and socio-economic variables*. Berlin: Kluwer Academic / Plenum Publishers.

Hout, M., & DiPrete, T. A. (2006). What we have learned: RC28s contributions to knowledge about social stratification. *Research into Social Stratification and Mobility, 24*, 1-20.

ILO. (1969). *International standard classification of occupations : Revised edition 1968*. New York: International Labour Office.

ILO. (1990). *ISCO-88 : International standard classification of occupations*. New York: International Labour Office.

Lambert, P. S. (2002). Handling occupational information. *Building Research Capacity, 4*, 9-12.

Lambert, P. S., & Tan, K. L. T. (2007). *Instructions for using the GEODE Portal, edition 0.3*. Stirling: GEODE Project Technical Paper No. 1, University of Stirling, and http://www.geode.stir.ac.uk

Lambert, P. S., Tan, K. L. T., Gayle, V., Prandy, K., & Turner, K. J. (2008 forthcoming). The importance of specificity in occupation-based social classifications. *International Journal of Sociology and Social Policy*.

Marsh, C. (1986). Occupationally based measures. In A. Jacoby (Ed.), *The measurement of social class* (pp. 1-47). London: Social Research Association.

McKnight, A., & Elias, P. (1997). A database of information on unit groups of the Standard Occupational Classification. In D. Rose & K. O'Reilly (Eds.), *Constructing classes* (pp. 116-145). Colchester, UK: University of Essex.

Modood, T. (2005). The educational attainments of ethnic minorities in Britain. In G. C. Loury, T. Modood & S. M. Teles (Eds.), *Ethnicity, social mobility and public policy: Comparing the US and UK* (pp. 288-308). Cambridge: Cambridge University Press.

ONS. (2000). *Standard occupational classification 2000, volume 1: Structure and description of unit groups*. London: Office for National Statistics.

ONS. (2003). *UK standard industrial classification of economic activities 2003*. London: Office for National Statistics.

OPCS. (1980). *Classification of occupations 1980*. London: Office for Population Censuses and Surveys.

OPCS. (1990). *Standard occupational classification, volume 1: Structure and definition of major, minor and unit groups*. London: Office for Population Censuses and Surveys.

Platt, L. (2005). *Migration and social mobility: The life chances of Britain's minority ethnic communities*. Bristol, UK: The Policy Press.

Reid, I. (1998). *Class in Britain*. London: Polity.

Rose, D., & Pevalin, D. J. (Eds.). (2003). *A researcher's guide to the National Statistics Socio-economic Classification*. London: Sage.

Routh, G. (1980). *Occupation and pay in Great Britain, 1906-79*. London: MacMillan.

Szreter, S. R. S. (1984). The genesis of the Registrar-General : social classification of occupations. *British Journal of Sociology, 35*(4), 522-546.

Tan, K. L. T., Gayle, V., Lambert, P. S., Sinnott, R. O., & Turner, K. J. (2006). GEODE- Sharing occupational data through the Grid. In S. J. Cox (Ed.), *Proceedings of the 5th UK e-Science All Hands Meeting* (pp. 534-541). Edinburgh: National e-Science Centre.

van Leeuwen, M. H. D., Maas, I., & Miles, A. (Eds.). (2005). *Marriage choices and class boundaries: Social endogamy in history*. Cambridge: Cambridge University Press.

Wright, E. O. (1985). *Classes*. London: Verso.

Wright, E. O. (Ed.). (2005). *Approaches to class analysis*. Cambridge: Cambridge University Press.

Author/s:
Lambert, Paul; Gayle, Vernon; Tan, Larry; Turner, Ken; SINNOTT, RICHARD; Prandy, Ken

Title:
Data curation standards and social science occupational information resources

Date:
2007

Citation:
Lambert, P., Gayle, V., Tan, L., Turner, K., Sinnott, R., & Prandy, K. (2007). Data curation standards and social science occupational information resources. International Journal of Digital Curation, 2(1), 73-91.

Publication Status:
Published

Persistent Link:
http://hdl.handle.net/11343/28826

File Description:
Data curation standards and social science occupational information resources